6L-04

情報格差多地点観測ヘテロデータ共有システムにおけるデータ取得手順

Data acquisition procedure for digital divide studies with heterogeneous data from multiple observation sites

岩田 翔汰 † 中平 勝子 † 北島 宗雄 †
Shota Iwata Katsuko T. Nakahira Muneo Kitajima

1 はじめに

インターネット利用状況の観測は，情報格差の分析の精度向上や，情報格差に関する新たな指標の開発に繋がることが期待される．また，UNESCO は Lena Resolution の中で，言語多様性の保護を訴え，サイバー空間へのユニバーサルアクセスを推奨し，インターネットにおける言語多様性のモニタリングを提案している [1]．この実現には，インターネット全体を網羅した観測が重要になる．観測したデータの共有の仕様を整えることで，インターネット全体を網羅する情報格差の観測が可能になるのではないかと考えられる [2]．インターネット全体を網羅する情報格差の観測は，複数の機関による観測が不可欠であるが，各機関の観測方針によって指標用データが異なることが予想され，データ量やその均質性に問題が生じる．これらの問題を解決するために，各機関同士の情報格差観測データの円滑な共有 · 相互利用が実現される必要がある．

伊藤らは，情報格差分析に利用するクロールデータや統計情報などを 1 次処理したデータを研究者間で流通させるために必要なデータ形式 Information Trade Handling Format（以下，ITHF）および，情報格差を分析するシステムを提案している [3, 4, 5, 6]．また，筆者らは，ITHF を流通させる基盤の設計を提案している [7]．この中で，ITHF の一部のヘッダと新たに設定した項目をデータ所在情報カタログ（データ所在地の URI，データの種類，対象国名等が登録される）へ集約 · 管理する方法を提案している．本稿では，情報格差観測データである ITHF ファイルの共有システムの設計と，共有システムから観測データへアクセスし，データを取得する手順の設計を行う．

2 情報格差観測データの処理 · 利用過程

図 1 は情報格差観測データの処理 · 利用過程を示している．縦方向はデータの加工を表し，横方向はデータの入出力を表している．時間の経過によってデータの状態が変化する．

0 次データ：インターネット利用状況の観測は，公的機関が IT 機器の普及率や通信網の整備状況等の統計を公開することや，研究機関が Web ページのクローリングによって行われる．クローリングを行うことで，URL やリンク情報，コンテンツ等の様々な情報が得られる．そして，情報格差の観測に直接使用しないデータを取り除くクレンジング処理を施し，また，地理情報の取得や言語解析を行うことで，URL データ，LINK データ，サーバ位置情報，言語解析データといった情報格差データが取得できる．また，公的機関が発行する統計データも情報格差観測に用いるため，分析に適した形に成型する．これらのデータの取得や各種の処理は各機関によって行われる．

1 次データ：情報格差観測データを公開するため，0 次データ（生データ）から公開用のデータフォーマットである ITHF ファイルが作成される．そして，各機関が個々に設置するサーバで ITHF ファイルへのアクセスが可能な状態で保管される．また，その ITHF ファイルへアクセスする手段となる URL や ITHF ファイルの概要データを提供する集中管理型のデータ所在カタログが設置される．

2 次データ：サーバ上に公開された 1 次データを取得し，分析するために必要なモジュール等を使用して作成される．分析し得られた結果が公開されることで，分析モジュールの開発が促進され，それに伴い，1 次データである情報格差観測データの公開が促進されることが期待される．

3 多地点ヘテロデータ

これまでの情報格差に関する研究は，各機関が独自に収集したデータや，公的機関によって公開されたデータを基に分析が行われてきた．このため，機関によって保持しているデータに偏りが存在する．各機関がデータを共有し，それぞれのデータを統合的に分析することで情報格差の分析精度が向上することが期待される．また，情報格差の分析に Web ページのクローリング結果を用いる場合，各機関が行えるクローリングには限りがあり，インターネットの一部分しか行えない．このため，各機関が行った部分的なクローリング結果を共有することで，インターネット全体の実態を反映する情報格差分析結果を導くことにつながるのではないかと考えられる．

また，データは各機関独自のフォーマットで管理されている状況下では，収集したデータやデータから得られた分析結果を他の機関（研究者）が再利用することは困難であった．情報格差について研究を行う者や，Web ページのクローリングを行う企業，各種統計調査を行う機関等が，所有しているデータを情報格差分析に使用することができる ITHF ファイルに変換することで，多種類の情報を同一の方法で分析することが可能になる．また，ITHF ファイルの所在情報カタログが集約されることで，これまでに様々な地点で取得された多くの情報について分析を行うことが容易になる．

4 共有システム

本システムで共有されるデータは，ITHF ファイルのヘッダの一部（以下，ITHF 概要データ）と ITHF ファイルが所在す

† 長岡技術科学大学

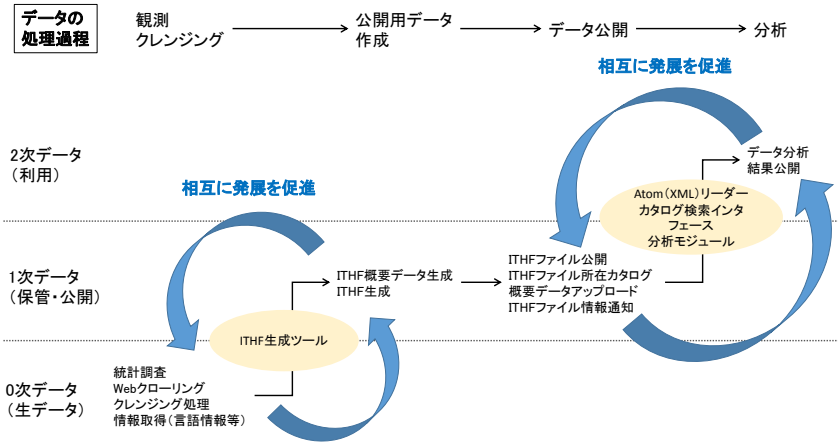

図 1　情報格差観測データの処理・利用過程

る URL である．共有システムは，主に，ITHF 概要データアップローダ，ITHF ファイル所在カタログ，カタログ内を検索するインタフェース，カタログに新規登録された ITHF ファイルの情報を通知する機能の 4 要素と，各機関が ITHF ファイルを公開するサーバ群で構成される．

ITHF の作成：各機関が独自に所有するデータから ITHF ファイルを作成する際に，同時に ITHF 概要データの生成も行う．ITHF 概要データには，作成した ITHF ファイルの対象国名，ITHF ファイルサイズ，ITHF 作成日時，ITHF 作成者名，ITHF 所有組織，生データ作成日時，生データ作成者名，生データ所有組織，データ種類を記述する．共有システムの設計に際し，伊藤らの提案した ITHF ヘッダに生データの作成日時や作成者，所有組織を追記する形で改定されている．

ITHF の公開：ITHF ファイル作成者は，ITHF 概要データと公開する ITHF ファイルの URI を合わせてアップローダにアップロードする．ここでアップロードされた情報は自動的に ITHF ファイル所在情報カタログとして格納される．同時に，アップロードされた情報から Atom 形式を拡張した XML 形式のデータを生成し，カタログに新規登録された ITHF ファイルの情報を提供する．

ITHF の取得：ITHF ファイルを情報格差の分析に使う研究者は，あらかじめ XML ファイルを Atom（XML）リーダに登録しておくことで，新たに公開された ITHF ファイルの概要データや ITHF ファイルへのリンク URI を入手できる．また，データベースの検索インタフェースを使用し，国名や日時，データ種類を指定することで，ITHF ファイルの概要データやリンク URI を入手可能にする．これらの方法で得られた URI に存在するファイルをダウンロードすることで，ITHF ファイルを入手することが可能になる．ITHF ファイルを暗号化して公開し，共有システムの利用を登録制にすることで，データの改ざんや再配布を防止する．この時，ダウンロードに用いる IP アドレスも登録することで，不正にダウンロードされることも防ぐことを想定している．また，カタログの追加や検索を高速に行うため，ITHF ファイル所在情報カタログを，ITHF ファイルが対象とする国名で地理的に分割し，分散的に管理する工夫を行う．また，ITHF ファイルのダウンロードを高速化するため，ITHF ファイルをサーバ上で分割してダウンロード用ファイルを生成するようにようにサーバ群を構成する．

5　まとめ

本稿では，多地点で取得されたヘテロな情報格差観測データを共有するシステムの設計と，そのシステムからデータを取得する方法の設計を行った．この提案が，情報格差観測データの共有促進につながることを期待する．

今後は，これらの実装を行い，情報格差に関する研究を発展させるために，システムの改良を行っていきたい．

参考文献

[1] UNESCO：INTERNATUONAL MEETINGS ON MULTILINGUALISM IN CYBERSPACE 2008-2014 FINAL DOCUMENTS

[2] 三上喜貴，中平勝子，児玉茂昭：言語天文台からみた世界の情報格差，慶應義塾大学出版会，2014

[3] 伊藤公，中平勝子，三上喜貴：デジタル・デバイド研究/分析用 DB の構築，電子情報通信学会信越支部大会，p.177，2013

[4] 伊藤公，中平勝子，三上喜貴：国別ドメイン利活用分析のためのプロビジョンスキーム，情報処理学会第 76 回全国大会，第 1 分冊，pp.379-380，2014

[5] 伊藤公，中平勝子，三上喜貴：クロールデータのプロビジョンスキームにおけるファイル入出力機構の検証，第 13 回情報科学技術フォーラム，第 2 分冊，pp.135-136，2014

[6] 伊藤公，中平勝子，三上喜貴：大容量データ流通のためのファイルシステムの開発・評価，情報処理学会第 76 回全国大会，第 4 分冊，pp.805-806，2015

[7] 岩田翔汰，中平勝子，北島宗雄：サービス利用状況に着目した情報格差観測データ流通基盤の設計，第 14 回情報科学技術フォーラム，第 2 分冊，pp.171-174，2015